

Node discovery problem for a social network

Yoshiharu Maeno*
Social Design Group

Abstract

Methods to solve a node discovery problem for a social network are presented. Covert nodes refer to the nodes which are not observable directly. They transmit the influence and affect the resulting collaborative activities among the persons in a social network, but do not appear in the surveillance logs which record the participants of the collaborative activities. Discovering the covert nodes is identifying the suspicious logs where the covert nodes would appear if the covert nodes became overt. The performance of the methods is demonstrated with a test dataset generated from computationally synthesized networks and a real organization.

*Yoshiharu Maeno, Ph.D. is a founder management consultant of Social Design Group, Sengoku 1-6-38F, Bunkyo-ku, Tokyo 112-0011 Japan. Telephone: +81-90-8009-1968. Email: maeno.yoshiharu@socialdesigngroup.com.

1 Introduction

Covert nodes refer to persons who transmit the influence and affect the resulting collaborative activities among the persons in a social network, but do not appear in the surveillance logs which record the participants of the activities. The covert nodes are not observable directly. It aids us in discovering and approaching to the covert nodes to identify the suspicious surveillance logs where the covert nodes would appear if they became overt. I call this problem a node discovery problem for a social network.

Where do we encounter such a problem? Globally networked clandestine organizations such as terrorists, criminals, or drug smugglers are great threat to the civilized societies [Sageman (2004)]. Terrorism attacks cause great economic, social and environmental damage. Active non-routine responses to the attacks are necessary as well as the damage recovery management. The short-term target of the responses is the arrest of the perpetrators. The long-term target of the responses is identifying and dismantling the covert organizational foundation which raises, encourages, and helps the perpetrators. The threat will be mitigated and eliminated by discovering covert leaders and critical conspirators of the clandestine organizations. The difficulty of such discovery lies in the limited capability of surveillance. Information on the leaders and critical conspirators are missing because it is usually hidden by the organization intentionally.

Let me show an example in the 9/11 terrorist attack in 2001 [Krebs (2002)]. Mustafa A. Al-Hisawi, whose alternate name was Mustafa Al-Hawsawi, was alleged to be a wire-puller who had acted as a financial manager of Al Qaeda. He had attempted to help terrorists enter the United States, and provided the hijackers of the 4 aircrafts with financial support worth more than 300,000 dollars. Furthermore, Osama bin Laden is suspected to be a wire-puller behind Mustafa A. Al-Hisawi and the conspirators behind the hijackers. These persons were not recognized as wire-pullers at the time of the attack. They were the nodes to discover from the information on the collaborative activities of the perpetrators and conspirators known at that moment.

In this paper, I present two methods to solve the node discovery problem. One is a heuristic method in [Maeno (2009)], which demonstrates a simulation experiment of the node discovery problem for the social network of the 9/11 perpetrators. The other is a statistical inference method which I propose in this paper. The method employs the maximal likelihood estimation and an anomaly detection technique. Section 3 defines the node discovery problem mathematically. Section 4 presents the two methods. Section 5 introduces the test dataset generated from computationally synthesized networks and a real clandestine organization. Section 6 demonstrates the performance characteristics

of the methods (precision, recall, and van Rijsbergen's F measure [Korfhage (1997)]). Section 7 presents the issues and future perspectives as concluding remarks. Section 2 summarizes the related works.

2 Related Work

The social network analysis is a study of social structures made of nodes which are linked by one or more specific types of relationship. Examples of the relationship are influence transmission in communication or presence of trust in collaboration [Lavrač (2007)]. Network topological characteristics of clandestine terrorist organizations [Krebs (2002)] and criminal organizations [Klerks (2002)] are studied. Trade-off between staying secret and efficiently securing coordination and control is of particular interest [Morselli (2007)]. The impact of the network topology to the trade-off is analyzed [Lindelauf (2009)].

Research interests have been moving from describing organizational structure to discovering dynamical phenomena on a social network. A link discovery predicts the existence of an unknown link between two nodes from the information on the known attributes of the nodes and the known links [Clauset (2008)]. It is one of the tasks of link mining [Getoor (2005)]. The link discovery techniques are combined with domain-specific heuristics. The collaboration between scientists can be predicted from the published co-authorship [Liben-Nowell (2004)]. The friendship between people is inferred from the information available on their web pages [Adamic (2003)].

Markov random network is a model of the joint probability distribution of random variables. It is an undirected graphical model similar to a Bayesian network. The Markov random network is used to learn the dependency between the links which shares a node. The Markov random network is one of the dependence graphs [Frank (1986)], which models the dependency between links. Extension to hierarchical models [Lazega (1999)], multiple networks (treating different types of relationships) [Pattison (1999)], valued networks (with nodal attributes) [Robins (1999)], higher order dependency between the links which share no nodes [Pattison (2002)], and 2-block chain graphs (associating one set of explanatory variables with the other set of outcome variables) [Robins (2001)] are studied. A family of such extensions and model elaborations is named the exponential random graph [Anderson (1999)].

In addition to the link discovery, the related research topics are the exploration of an unknown network structure [Newman (2007)], the discovery of a community structure [Palla (2005)], the inference of a network topology [Rabbat (2008)], the detection of an anomaly in a network [Silva (2009)], and the discovery of unknown nodes [Maeno (2007)], [Maeno (2009)]. Stochas-

tic modeling to predict terrorism attacks [Singh (2004)] is relevant practically. The idea of machine learning of latent variables [Silva (2006)] is potentially applicable to discovering an unknown network structure.

3 Problem definition

The node discovery problem is defined mathematically in this section. A node represents a person in a social network. A link represents a relationship which transmits the influence between persons. The symbols n_j ($j = 0, 1, \dots$) represent the nodes. Some nodes are overt (observable), but the others are covert (unobservable). \mathbf{O} denotes the overt nodes; $\{n_0, n_1, \dots, n_{N-1}\}$. Its cardinality is $|\mathbf{O}| = N$. $\mathbf{C} = \overline{\mathbf{O}}$ denotes the covert nodes; $\{n_N, n_{N+1}, \dots, n_{M-1}\}$. Its cardinality is $|\mathbf{C}| = M - N$. The whole nodes in a social network is $\mathbf{O} \cup \mathbf{C}$. The number of the nodes is M . The unobservability of the covert nodes arises either from a technical defect of surveillance means or an intentional cover-up operation.

The symbol δ_i represent a set of participants in a particular collaborative activity. It is the i -th activity pattern among the nodes. A pattern δ_i is a set of nodes; δ_i is a subset of $\mathbf{O} \cup \mathbf{C}$. For example, the nodes in an collaborative activity pattern are those who joined a particular conference call. That is, a pattern is a co-occurrence among the nodes [Rabbat (2008)]. The unobservability of the covert nodes does not affect the activity patterns themselves.

A simple hub-and-spoke model is assumed as a model of the influence transmission over the links resulting the collaborative activities among the nodes. The way how the influence is transmitted governs the set of possible activity patterns $\{\delta_i\}$. The network topology and the influence transmission are described by some probability parameters. The probability where the influence transmits from an initiating node n_j to a responder node n_k is r_{jk} . The influence transmits to multiple responders independently in parallel. It is similar to the degree of collaboration probability in trust modeling [Lavrač (2007)]. The constraints are $0 \leq r_{jk}$ and $\sum_{k \neq j} r_{jk} \leq 1$. The quantity f_j is the probability where the node n_j becomes an initiator. The constraints are $0 \leq f_j$ and $\sum_{j=0}^{N-1} f_j = 1$. These parameters are defined for the whole nodes in a social network (both the nodes in \mathbf{O} and \mathbf{C}).

A surveillance log d_i records a set of the overt nodes in a collaborative activity pattern; δ_i . It is given by eq.(1). A log d_i is a subset of \mathbf{O} . The number of data is D . A set $\{d_i\}$ is the whole surveillance logs dataset.

$$d_i = \delta_i \cap \mathbf{O} \quad (0 \leq i < D). \quad (1)$$

Note that neither an individual node nor a single link alone can be observed directly, but nodes can be observed collectively as a collaborative activity pattern. The dataset $\{d_i\}$ can be expressed by a 2-dimensional

$D \times N$ matrix of binary variables \mathbf{d} . The presence or absence of the node n_j in the data d_i is indicated by the elements in eq.(2).

$$d_{ij} = \begin{cases} 1 & \text{if } n_j \in d_i \\ 0 & \text{otherwise} \end{cases} \quad (0 \leq i < D, 0 \leq j < N). \quad (2)$$

Solving the node discovery problem means identifying all the surveillance logs where covert nodes would appear if they became overt. In other words, it means to identifying the logs for which $d_i \neq \delta_i$ holds because of the covert nodes belonging to \mathbf{C} .

4 Solution

4.1 Heuristic method

A heuristic method to solve the node discovery problem is studied in [Maeno (2009)]. The method is reviewed briefly.

At first, every node which appears in the dataset $\{d_i\}$ is classified into one of the clusters c_l ($0 \leq l < C$). The number of clusters is C , which depends on the prior knowledge. Mutually close nodes form a cluster. The measure of closeness between a pair of nodes is evaluated by the Jaccard's coefficient [Liben-Nowell (2004)]. It is used widely in link discovery, web mining, or text processing. The Jaccard's coefficient between the nodes n and n' is defined by eq.(3). The function $B(s)$ in eq.(3) is a Boolean function which returns 1 if the proposition s is trueCor 0 otherwise. The operators \wedge and \vee are logical AND and OR.

$$J(n, n') = \frac{\sum_{i=0}^{D-1} B(n \in d_i \wedge n' \in d_i)}{\sum_{i=0}^{D-1} B(n \in d_i \vee n' \in d_i)}. \quad (3)$$

The k-medoids clustering algorithm [Hastie (2001)] is employed for classification of the nodes. It is an EM (expectation-maximization) algorithm similar to the k-means algorithm for numerical data. A medoid node locates most centrally within a cluster. It corresponds to the center of gravity in the k-means algorithm. The clusters and the modoid nodes are re-calculated iteratively until they converge into a stable structure. The k-medoids clustering algorithm may be substituted by other clustering algorithms such as hierarchical clustering or self-organizing mapping.

Then, suspiciousness of every surveillance log d_i as a candidate where the covert nodes would appear is evaluated with a ranking function $s(d_i)$. The ranking function returns higher value for a more suspicious log. The strength of the correlation between the log d_i and the cluster c_l is defined by $w(d_i, c_l)$ in eq.(4) as a preparation.

$$w(d_i, c_l) = \max_{n_j \in c_l} \frac{B(n_j \in d_i)}{\sum_{i=0}^{D-1} B(n_j \in d_i)}. \quad (4)$$

The ranking function takes $w(d_i, c_l)$ as an input. Various forms of ranking functions can be constructed. For example, [Maeno (2009)] studied a simple form in eq.(5) where the function $u(x)$ returns 1 if the real variable x is positive, or 0 otherwise.

$$\begin{aligned} s(d_i) &\propto \sum_{l=0}^{C-1} u(w(d_i, c_l)) \\ &= \sum_{l=0}^{C-1} B(d_i \cap c_l \neq \phi). \end{aligned} \quad (5)$$

The i -th most suspicious log is given by $d_{\sigma(i)}$ where $\sigma(i)$ is calculated by eq.(6). Suspiciousness $s(d_{\sigma(i)})$ is always larger than $s(d_{\sigma(i')})$ for any $i < i'$.

$$\sigma(i) = \arg \max_{m \neq \sigma(n) \text{ for } \forall n < i} s(d_m) \quad (1 \leq i \leq D). \quad (6)$$

The computational burden of the method remains light as the number of nodes and surveillance logs increases. The method is expected to work generally for clustered networks but moderately even if the network topological and stochastic mechanism to generate the surveillance logs is not understood well. The method works without the knowledge about the hub-and-spoke model; the parametric form with r_{jk} and f_j in Section 3. The result, however, can not be very accurate because of the heuristic nature. A statistical inference method which requires heavy computational burden, but outputs more accurate results is presented next.

4.2 Statistical inference method

The statistical inference method employs the maximal likelihood estimation to infer the topology of the network, and applies an anomaly detection technique to retrieve the suspicious surveillance logs which are not likely to realize without the covert nodes. The maximal likelihood estimation is a basic statistical method used for fitting a statistical model to data and for providing estimates for the model's parameters. The anomaly detection refers to detecting patterns in a given dataset that do not conform to an established normal behavior.

A single symbol θ represent both of the parameters r_{jk} and f_j for the nodes in \mathcal{O} . θ is the target variable, the value of which needs to be inferred from the surveillance log dataset. The logarithmic likelihood function [Hastie (2001)] is defined by eq.(7). The quantity $p(\{d_i\}|\theta)$ denote the probability where the surveillance log dataset $\{d_i\}$ realizes under a given θ .

$$L(\theta) = \log(p(\{d_i\}|\theta)). \quad (7)$$

The individual surveillance logs are assumed to be independent. eq.(7) becomes eq.(8).

$$L(\theta) = \log\left(\prod_{i=0}^{D-1} p(d_i|\theta)\right)$$

$$= \sum_{i=0}^{D-1} \log(p(d_i|\theta)). \quad (8)$$

The quantity $q_{i|jk}$ in eq.(9) is the probability where the presence or absence of the node n_k as a responder to the stimulating node n_j coincides with the surveillance log d_i .

$$q_{i|jk} = \begin{cases} r_{jk} & \text{if } d_{ik} = 1 \text{ for given } i \text{ and } j \\ 1 - r_{jk} & \text{otherwise} \end{cases} \quad (9)$$

Eq.(9) is equivalent to eq.(10) since the value of d_{ik} is either 0 or 1.

$$q_{i|jk} = d_{ik}r_{jk} + (1 - d_{ik})(1 - r_{jk}). \quad (10)$$

The probability $p(\{d_i\}|\theta)$ in eq.(8) is expressed by eq.(11).

$$p(d_i|\theta) = \sum_{j=0}^{N-1} d_{ij} f_j \prod_{0 \leq k < N \wedge k \neq j} q_{i|jk}. \quad (11)$$

The logarithmic likelihood function takes an explicit formula in eq.(12). The case $k = j$ in multiplication (\prod_k) is included since $d_{ik}^2 = d_{ik}$ always holds.

$$\begin{aligned} L(\theta) = \sum_{i=0}^{D-1} \log & \left(\sum_{j=0}^{N-1} d_{ij} f_j \prod_{k=0}^{N-1} \{1 - d_{ik} \right. \\ & \left. + (2d_{ik} - 1)r_{jk}\} \right). \end{aligned} \quad (12)$$

The maximal likelihood estimator $\hat{\theta}$ is obtained by solving eq.(13). It gives the values of the parameters r_{jk} and f_j . A pair of nodes n_j and n_k for which $r_{jk} > 0$ possesses a link between them.

$$\hat{\theta} = \arg \max_{\theta} L(\theta). \quad (13)$$

A simple incremental optimization technique; the hill climbing method (or the method of steepest descent) is employed to solve eq.(13). Non-deterministic methods such as simulated annealing [Hastie (2001)] can be employed to strengthen the search ability and to avoid sub-optimal solutions. These methods search more optimal parameter values around the present values and update them as in eq.(14) until the values converge.

$$\begin{cases} r_{jk} \rightarrow r_{jk} + \Delta r_{jk} \\ f_j \rightarrow f_j + \Delta f_j \end{cases} \quad (0 \leq j, k < N). \quad (14)$$

The change in the logarithmic likelihood function can be calculated as a product of the derivatives (differential coefficients with regard to r and f) and the amount of the updates in eq.(15). The update Δr_{nm} and Δf_n should be in the direction of the steepest ascent in the landscape of the logarithmic likelihood function.

$$\Delta L(\theta) = \sum_{n,m=0}^{N-1} \frac{\partial L(\theta)}{\partial r_{nm}} \Delta r_{nm} + \sum_{n=0}^{N-1} \frac{\partial L(\theta)}{\partial f_n} \Delta f_n. \quad (15)$$

The derivatives with regard to r are given by eq.(16).

$$\begin{aligned}
\frac{\partial L(\boldsymbol{\theta})}{\partial r_{nm}} &= \sum_{i=0}^{D-1} [f_n \mathbf{d}_{in} (2\mathbf{d}_{im} - 1)] \\
&\times \prod_{k \neq m} \{1 - d_{ik} + (2\mathbf{d}_{ik} - 1)r_{nk}\} \\
&\div \sum_{j=0}^{N-1} \mathbf{d}_{ij} f_j \prod_{k=0}^{N-1} \{1 - \mathbf{d}_{ik} + (2\mathbf{d}_{ik} - 1)r_{jk}\}.
\end{aligned} \tag{16}$$

The derivatives with regard to f are given by eq.(17).

$$\begin{aligned}
\frac{\partial L(\boldsymbol{\theta})}{\partial f_n} &= \sum_{i=0}^{D-1} [\mathbf{d}_{in} \prod_{k=0}^{N-1} \{1 - \mathbf{d}_{ik} + (2\mathbf{d}_{ik} - 1)r_{nk}\}] \\
&\div \sum_{j=0}^{N-1} \mathbf{d}_{ij} f_j \prod_{k=0}^{N-1} \{1 - \mathbf{d}_{ik} + (2\mathbf{d}_{ik} - 1)r_{jk}\}.
\end{aligned} \tag{17}$$

The ranking function $s(d_i)$ is the inverse of the probability at which d_i realizes under the maximal likelihood estimator $\hat{\boldsymbol{\theta}}$. According to the anomaly detection technique, it gives a higher return value to the suspicious surveillance logs which are less likely to realize without the covert nodes. The ranking function is given by eq.(18).

$$s(d_i) = \frac{1}{p(d_i | \hat{\boldsymbol{\theta}})}. \tag{18}$$

The i -th most suspicious log is given by $d_{\sigma(i)}$ by the same formula in eq.(6).

5 Test Dataset

5.1 Network

Two classes of networks are employed to generate a test dataset for performance evaluation of the two methods. The first class is computationally synthesized networks. The second class is a real clandestine organization.

The networks [A] in Figure 1 and [B] in Figure 2 are synthesized computationally. They are based on the Barabási-Albert model [Barabási (1999)] with a cluster structure. The Barabási-Albert model grows with preferential attachment. The probability where a newly coming node n_k connects a link to an existing node n_j is proportional to the nodal degree of n_j ($p(k \rightarrow j) \propto K(n_j)$). The occurrence frequency of the nodal degree tends to be scale-free ($F(K) \propto K^a$). In the Barabási-Albert model with a cluster structure, every node n_j is assigned a pre-determined cluster attribute $c(n_j)$ to which it belongs. The number of clusters is C . The probability $p(k \rightarrow j)$ is modified to eq.(19). cluster contrast parameter η is introduced. Links between the

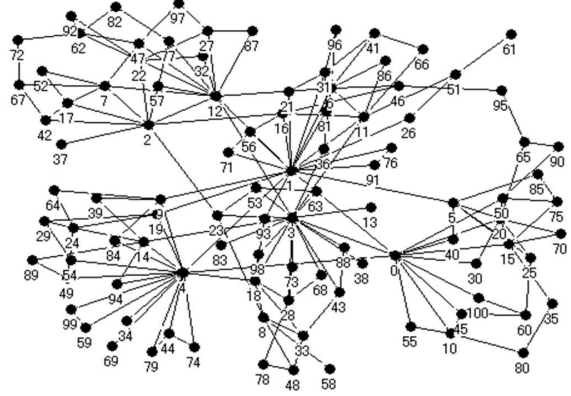


Figure 1: Computationally synthesized network [A] which consists of 101 nodes and 5 clusters. Cluster contrast parameter is $\eta = 50$. The network is relatively more clustered. The node n_{12} is a typical hub node. The node n_{75} is a typical peripheral node.

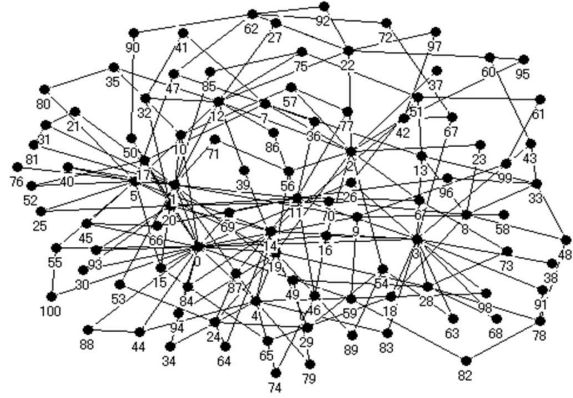


Figure 2: Computationally synthesized network [B] which consists of 101 nodes and 5 clusters. Cluster contrast parameter is $\eta = 2.5$. The network is relatively less clustered. The node n_{12} is a typical hub node. The node n_{48} is a typical peripheral node.

clusters appear less frequently as η increases. The initial links between the clusters are connected at random before growth by preferential attachment starts.

$$p(k \rightarrow j) \propto \begin{cases} \eta(C-1)K(n_j) & \text{if } c(n_j) = c(n_k) \\ K(n_j) & \text{otherwise} \end{cases} \quad (19)$$

Hub nodes are those which have a nodal degree larger than the average. The node n_{12} in the network [A] in Figure 1 is a typical hub node. Peripheral nodes are those which have a nodal degree smaller than the average. The node n_{75} in the network [A] in Figure 1 is a typical peripheral node.

The network in Figure 3 represents a real clandestine organization. It is a global mujahedin organization which was analyzed in [Sageman (2004)]. The mujahedin in the global Salafi jihad means Muslim fighters in Salafism (Sunni Islamic school of thought) who struggle to establish justice on earth. Note that jihad does not necessarily refer to military exertion. The organization consists of 107 persons and 4 regional sub-networks. The sub-networks represent Central Staffs (n_{CSj}) including the node n_{ObL} , Core Arabs (n_{CAj}) from the Arabian Peninsula countries and Egypt, Maghreb Arabs (n_{MAj}) from the North African countries, and Southeast Asians (n_{SAj}). The network topology is not simply hierarchical. The 4 regional sub-networks are connected mutually in a complex manner.

The node representing Osama bin Laden; n_{ObL} is a hub ($K(n_{ObL}) = 8$). He is believed to be the founder of the organization, and said to be the covert leader who provides operational commanders in regional sub-networks with financial support in many terrorism attacks including 9/11 in 2001. His whereabouts are not known despite many efforts in investigation and capture.

The topological characteristics of the above mentioned networks are summarized in Table 1. The global mujahedin organization has a relatively large Gini coefficient of the nodal degree; $G = 0.35$ and a relatively large average clustering coefficient [Watts (1998)]; $\langle W(n_j) \rangle = 0.54$. In economics, the Gini coefficient is a measure of inequality of income distribution or of wealth distribution. A larger Gini coefficient indicates lower equality. The values mean that the organization possesses hubs and a cluster structure. The values also indicate that the computationally synthesized network [A] is more clustered and close to the global mujahedin organization while the network [B] is less clustered.

5.2 Test Dataset

The test dataset $\{d_i\}$ is generated from each network in 5.1 in the 2 steps below.

In the first step, the collaborative activity patterns $\{\delta_i\}$ are generated D times according to the influence

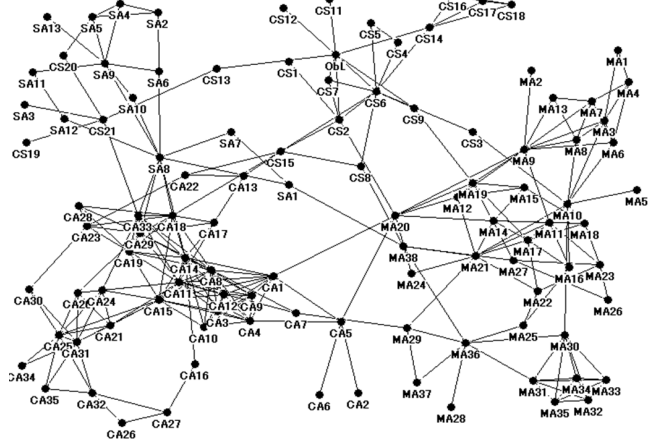


Figure 3: Social network representing a global mujahedin (Jihad fighters) organization [Sageman (2004)], which consists of 107 nodes and 4 regional sub-networks. The sub-networks represent Central Staffs (n_{CSj}) including the node n_{ObL} , Core Arabs (n_{CAj}), Maghreb Arabs (n_{MAj}), and Southeast Asians (n_{SAj}). The node n_{ObL} is Osama bin Laden who many believe is the founder of the organization.

Table 1: The number of nodes M , the number of clusters C , the average degree $\langle K(n_j) \rangle$, the average clustering coefficient $\langle W(n_j) \rangle$, and the Gini coefficient G of the computationally synthesized networks (CSN) [A] and [B], and the global mujahedin organization (GMO).

Model	M	C	η	$\langle K \rangle$	$\langle W \rangle$	G
CSN [A]	101	5	50	3.6	0.42	0.36
CSN [B]	101	5	2.5	3.9	0.22	0.37
GMO	107	-	-	5.1	0.54	0.35

transmission under the true value of θ . A pattern includes both an initiator node n_j and multiple responder nodes n_k . An example is $\delta_{\text{ex1}} = \{n_{\text{CS1}}, n_{\text{CS2}}, n_{\text{CS6}}, n_{\text{CS7}}, n_{\text{CS9}}, n_{\text{ObL}}, n_{\text{CS11}}, n_{\text{CS12}}, n_{\text{CS14}}\}$ for the global mujahedin organization in Figure 3.

In the second step, the surveillance log dataset $\{d_i\}$ is generated by deleting the covert nodes belonging to \mathcal{C} from the patterns $\{\delta_i\}$. The example δ_{ex1} results in the surveillance log $d_{\text{ex1}} = \delta_{\text{ex1}} \cap \overline{\mathcal{C}} = \{n_{\text{CS1}}, n_{\text{CS2}}, n_{\text{CS6}}, n_{\text{CS7}}, n_{\text{CS9}}, n_{\text{CS11}}, n_{\text{CS12}}, n_{\text{CS14}}\}$ if Osama bin Laden is a cover node; $\mathcal{C} = n_{\text{ObL}}$. The covert node in \mathcal{C} may appear multiple times in the collaborative activity patterns $\{\delta_i\}$. The number of the target logs to identify D_t is given by eq.(20).

$$D_t = \sum_{i=0}^{D-1} B(d_i \neq \delta_i). \quad (20)$$

In the performance evaluation in Section 6, a few assumptions are made for simplicity. The probability f_j does not depend on the nodes ($f_j = 1/M$). The value of the probability r_{jk} is either 1 when a link is present between nodes, or 1 otherwise. It means that the number of the possible collaborative activity patterns is bounded. The influence transmission is symmetrically bi-directional; $r_{jk} = r_{kj}$.

6 Performance

6.1 Performance measure

Three measures, precision, recall, and van Rijsbergen's F measure [Korfuge (1997)], are used to evaluate the performance of the methods. They are commonly used in information retrieval such as search, document classification, and query classification. The precision p is used as evaluation criteria, which is the fraction of the number of relevant data to the number of the all data retrieved by search. The recall r is the fraction of the number of the data retrieved by search to the number of the all relevant data. The relevant data refers to the data where $d_i \neq \delta_i$. They are given by eq.(21) and eq.(22) They are functions of the number of the retrieved data D_r . It can take the value from 1 to D . The data is retrieved in the order of $d_{\sigma(1)}, d_{\sigma(2)}, \dots, d_{\sigma(D_r)}$.

$$p(D_r) = \frac{\sum_{i=1}^{D_r} B(d_{\sigma(i)} \neq \delta_{\sigma(i)})}{D_r}. \quad (21)$$

$$r(D_r) = \frac{\sum_{i=1}^{D_r} B(d_{\sigma(i)} \neq \delta_{\sigma(i)})}{D_t}. \quad (22)$$

The F measure F is the harmonic mean of the precision and recall. It is given by eq.(23).

$$F(D_r) = \frac{1}{\frac{1}{2}(\frac{1}{p(D_r)} + \frac{1}{r(D_r)})}$$

$$= \frac{2p(D_r)r(D_r)}{p(D_r) + r(D_r)}. \quad (23)$$

The precision, recall, and F measure range from 0 to 1. All the measures take larger values as the performance of retrieval becomes better.

6.2 Comparison

The performance of the heuristic method and statistical inference method is compared with the test dataset generated from the computationally synthesized networks.

Figure 4 shows the precision $p(D_r)$ as a function of the rate of the retrieved data to the whole data D_r/D in case the hub node n_{12} in the computationally synthesized network [A] in Figure 1 is the target covert node to discover, $\mathcal{C} = \{n_{12}\}$. The three graphs are for [a] the statistical inference method, [b] the heuristic method ($C = 5$), and [c] the heuristic method ($C = 10$). The number of the surveillance logs in a test dataset is $D = 100$. The broken lines indicate the theoretical limit (the upper bound) and the random retrieval (the lower bound). The vertical solid line indicates the position where $D_r = D_t$. Figure 5 shows the recall $r(D_r)$ as a function of the rate D_r/D . Figure 6 shows the F measure $F(D_r)$ as a function of the rate D_r/D . The experimental conditions are the same as those for Figure 4. The performance of the heuristic method is moderately good if the number of clusters is known as prior knowledge. Otherwise, the performance would be worse. On the other hand, the statistical inference method surpasses the heuristic method and approaches to the theoretical limit.

Figure 7 shows the F measure $F(D_r)$ as a function of the rate D_r/D in case the hub node n_{12} in the network [B] in Figure 2 is the target covert node to discover. The two graphs are for [a] the statistical inference method and [b] the heuristic method ($C = 5$). The performance of the statistical inference method is still good while that of the heuristic method becomes worse in a less clustered network.

Figure 8 shows the F measure $F(D_r)$ as a function of the rate D_r/D in case the peripheral node n_{75} in the network [A] in Figure 1 is the target covert node to discover. Figure 9 shows the F measure $F(D_r)$ as a function of the rate D_r/D when the peripheral node n_{48} in the network [B] in Figure 2 is the target covert node to discover. The statistical inference method works fine while the heuristic method fails.

6.3 Application

I illustrate how the method aids the investigators in achieving the long-term target of the non-routine responses to the terrorism attacks. Let's assume that the investigators have surveillance logs of the members of the global mujahedin organization except Osama bin Laden by the time of the attack. Osama bin Laden

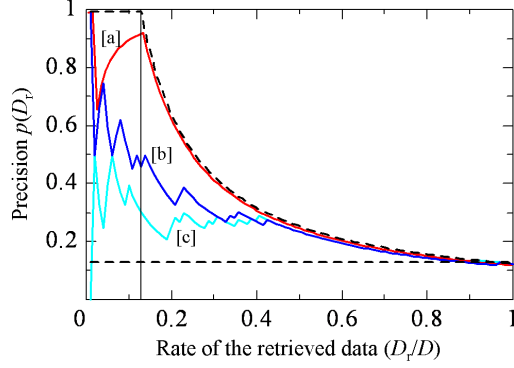


Figure 4: Precision $p(D_r)$ as a function of the rate of the retrieved data to the whole data D_r/D in case the hub node n_{12} in the computationally synthesized network [A] in Figure 1 is the target covert node to discover. $C = \{n_{12}\}$. $|C| = 1$. $|O| = 100$. $D = 100$. The three graphs are for [a] the statistical inference method, [b] the heuristic method ($C = 5$), and [c] the heuristic method ($C = 10$). The broken lines indicate the theoretical limit (the upper bound) and the random retrieval (the lower bound). The vertical solid line indicates the position where $D_r = D_t$.

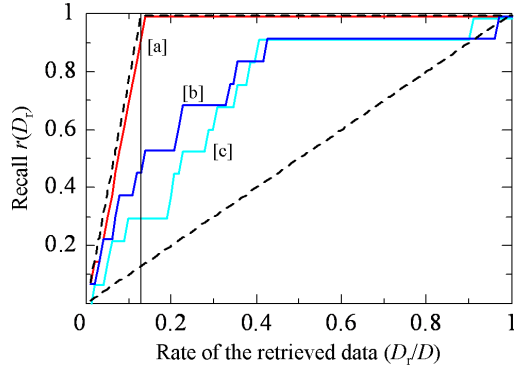


Figure 5: Recall $r(D_r)$ as a function of the rate D_r/D . The experimental conditions are the same as those for Figure 4.

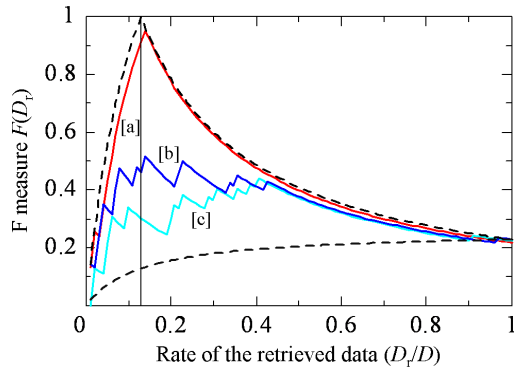


Figure 6: F measure $F(D_r)$ as a function of the rate D_r/D . The experimental conditions are the same as those for Figure 4.

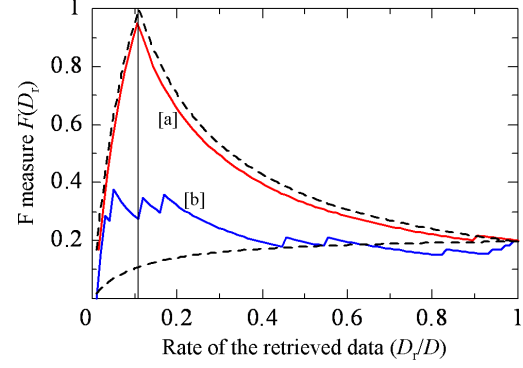


Figure 7: F measure $F(D_r)$ as a function of the rate D_r/D in case the hub node n_{12} in the computationally synthesized network [B] in Figure 2 is the target covert node to discover. Two graphs are for [a] the statistical inference method, and [b] the heuristic method ($C = 5$).

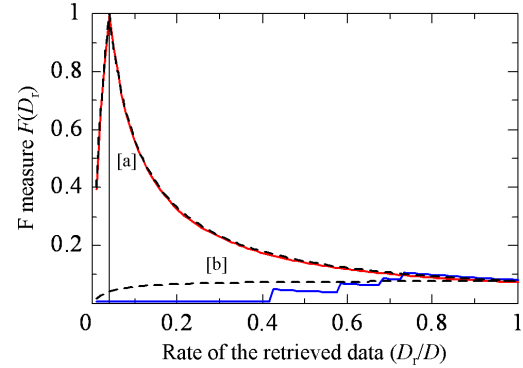


Figure 8: F measure $F(D_r)$ as a function of the rate D_r/D in case the peripheral node n_{75} in the computationally synthesized network [A] in Figure 1 is the target covert node to discover. Two graphs are for [a] the statistical inference method, and [b] the heuristic method ($C = 5$).

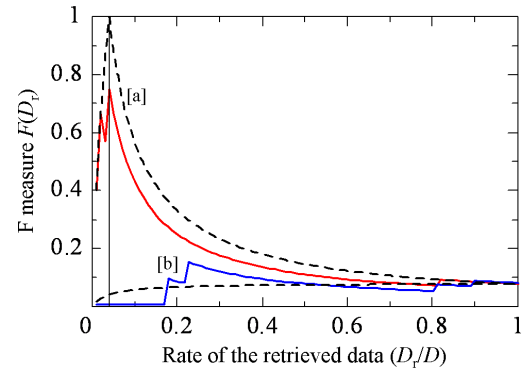


Figure 9: F measure $F(D_r)$ as a function of the rate D_r/D when the peripheral node n_{48} in the computationally synthesized network [B] in Figure 2 is the target covert node to discover. Two graphs are for [a] the statistical inference method, and [b] the heuristic method ($C = 5$).

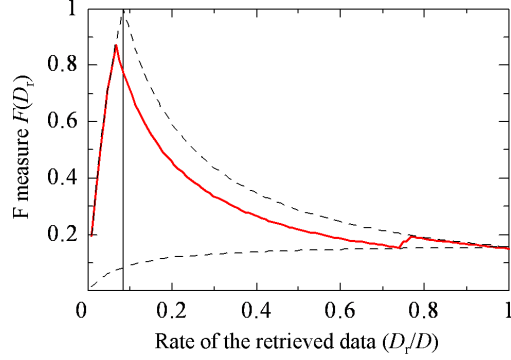


Figure 10: F measure $F(D_r)$ as a function of the rate of the retrieved data to the whole data D_r/D when the statistical inference method is applied in case the node n_{ObL} in Figure 3 is the target covert node to discover. $C = \{n_{ObL}\}$. $|C| = 1$. $|O| = 106$. The graph is for the statistical inference method. The broken lines indicate the theoretical limit and the random retrieval. The vertical solid line indicates the position where $D_r = D_t$.

does not appear in the logs. This is the assumption that the investigators neither know the presence of a wire-puller behind the attack nor recognize Osama bin Laden at the time of the attack.

The situation is simulated computationally like the problems addressed in 6.2. In this case, the node n_{ObL} in Figure 3 is the target covert node to discover, $C = \{n_{ObL}\}$. Figure 10 shows $F(D_r)$ as a function of the rate of the retrieved data to the whole data D_r/D when the statistical inference method is applied. The result is close to the theoretical limit. The most suspicious surveillance log $d_{\sigma(1)}$ includes all and only the neighbor nodes n_{CS1} , n_{CS2} , n_{CS6} , n_{CS7} , n_{CS9} , n_{CS11} , n_{CS12} , and n_{CS14} . This encourages the investigators to take an action to investigate an unknown wire-puller near these 8 neighbors; the most suspicious close associates. The investigators will decide to collect more detailed information on the suspicious neighbors. It may result in approaching to and finally capturing the covert wire-puller responsible for the attack.

The method, however, fails to identify two suspicious records $\delta_{A1} = \{n_{ObL}, n_{CS11}\}$ and $\delta_{A2} = \{n_{ObL}, n_{CS12}\}$. These nodes have a small nodal degree; $K(n_{CS11}) = 1$ and $K(n_{CS12}) = 1$. This shows that the surveillance logs on the nodes having small nodal degree do not provide the investigators with much clues for the covert nodes.

7 Conclusion

In this paper, I define the node discovery problem for a social network and present methods to solve the problem. The statistical inference method employs the maximal likelihood estimation to infer the topology of the

network, and applies an anomaly detection technique to retrieve the suspicious surveillance logs which are not likely to realize without the covert nodes. The precision, recall, and F measure characteristics are close to the theoretical limit for the discovery of the covert nodes in computationally synthesized networks and a real clandestine organization. In the investigation of a clandestine organization, the method aids the investigators in identifying the close associates and approaching to a covert leader or a critical conspirator.

The node discovery problem is encountered in many areas of business and social sciences. For example, in addition to the analysis of a clandestine organization, the method contributes to detecting an individual employee who transmits the influence to colleagues, but whose catalytic role is not recognized by company managers, may be critical in reorganizing a company structure.

I plan to address two issues for the future works. The first issue is to extend the hub-and-spoke model for the influence transmission. The model represents the radial transmission from an initiating node toward multiple responder nodes. Other types of influence transmission are present in many real social networks. Examples are serial chain-shaped influence transmission model or tree-like influence transmission model. The second issue is to develop a method to solve the variants of the node discovery problem. Discovering fake nodes, or spoofing nodes are also interesting problems to uncover the malicious intentions of the organization. A fake node is the person who does not exist in the organization, but appears in the surveillance. A spoofing node is the person who belongs to an organization, but appears as a different node in the surveillance logs.

References

- [Anderson (1999)] Anderson, C., Wasserman, S., and Crouch, B., 1999. A p^* primer, Logit models for social networks. *Social Networks* 21, 37-66.
- [Adamic (2003)] Adamic, L. A., Adar, E., 2003. Friends and neighbors on the web. *Social Networks* 25, 211-228.
- [Adamic (2001)] Adamic, L. A., Lukose, R. M., Puniyani, A. R., Huberman, B., 2001. Search in power-law networks. *Physical Review E* 64, 046135.
- [Barabási (1999)] Barabási, A. L., Albert, R., and Jeong, H., 1999. Mean-field theory for scale-free random networks. *Physica A* 272, 173-187.
- [Clauset (2008)] Clauset, A., Moore, C., and Newman, M. E. J., 2008. Hierarchical structure and the prediction of missing links in networks. *Nature* 453, 98-100.
- [Frank (1986)] Frank, O., and Strauss, D., 1986. Markov graphs. *Journal of the American Statistical Association* 81, 832-842.
- [Getoor (2005)] Getoor, L., and Diehl, C. P., 2005. Link mining: a survey. *ACM SIGKDD Explorations* 7, 3-12.
- [Hastie (2001)] Hastie, T., Tibshirani, R., Friedman, J., 2001. *The elements of statistical learning: Data mining, inference, and prediction* (Springer series in statistics). Springer-Verlag, Berlin.
- [Keila (2006)] Keila, P. S., Skillicorn, D. B., 2006. Structure in the Enron email dataset. *Journal of Computational & Mathematical Organization Theory* 11, 183-199.
- [Klerks (2002)] Klerks, P., 2002. The network paradigm applied to criminal organizations. *Connections* 24, 53-65.
- [Krebs (2002)] Krebs, V. E., 2002. Mapping networks of terrorist cells. *Connections* 24, 43-52.
- [Korfhuge (1997)] Korfhuge, R. R., 1997. *Information storage and retrieval*. Wiley.
- [Lavrač (2007)] Lavrač, N., Ljubič, P., Urbančič, T., Papa, G., Jermol, M., and Bollhalter, S., 2007. Trust modeling for social networks using reputation and collaboration estimates. *IEEE Transactions on Systems, Man, & Cybernetics Part C* 37, 429-439.
- [Lazega (1999)] Lazega, E. and Pattison, P., 1999. Multiplexity, generalized exchange and cooperation in organizations. *Social Networks* 21, 67-90.
- [Liben-Nowell (2004)] Liben-Nowell, D., and Kleinberg, J., 2004. The link prediction problem for social networks. *Journal of American Society of Information Science and Technology* 58, 1019-1031.
- [Lindelauf (2009)] Lindelauf, R., Borm, P., and Hamers, H., 2009. The influence of secrecy on the communication structure of covert networks. *Social Networks* 31, 126-137.
- [Maeno (2007)] Maeno, Y., and Ohsawa, Y., 2007. Human-computer interactive annealing for discovering invisible dark events. *IEEE Transactions on Industrial Electronics* 54, 1184-1192.
- [Maeno (2009)] Maeno, Y., and Ohsawa, Y., 2009. Analyzing covert social network foundation behind terrorism disaster. *International Journal of Services Sciences* 2, 125-141.
- [Morselli (2007)] Morselli, C., Giguère, C., Petit, K., 2007. The efficiency/security trade-off in criminal networks. *Social Networks* 29, 143-153.
- [Newman (2007)] Newman, M. E. J., and Leicht, E. A., 2007. Mixture models and exploratory analysis in networks. *Proceedings of the National Academy of Sciences USA* 104, 9564-9569.
- [Palla (2005)] Palla, G., Derényi, I., Farkas, I., and Vicsek, T., 2005. Uncovering the overlapping community structure of complex networks in nature and society. *Nature* 435, 814-818.
- [Pattison (2002)] Pattison, P., and Robins, G. L., 2005. Neighborhood-based models for social networks. *Sociological Methodology* 32, 301-337.

- [Pattison (1999)] Pattison, P., and Wasserman, S., 1999. Logit models and logistic regressions for social networks. *British Journal of Mathematical and Statistical Psychology* 52, 169-194.
- [Rabbat (2008)] Rabbat, M. G., Figueiredo, M. A. T., and Nowak, R. D., 2008. Network Inference from co-occurrences. *IEEE Transactions on Information Theory* 54, 4053-4068.
- [Robins (2001)] Robins, G. J., Pattison, P. E., and Elliott, P., 2001. Network models for social influence processes. *Psychometrika* 66, 161-190.
- [Robins (1999)] Robins, G. L., Pattison, P. E., and Wasserman, S., 1999. Logit models and logistic regressions for social networks. *Psychometrika* 64, 371-394.
- [Sageman (2004)] Sageman, M., 2004. Understanding terror networks. University of Pennsylvania Press.
- [Silva (2009)] Silva, J., and Willett, R., 2009. Hypergraph-based anomaly detection of high-dimensional co-occurrences. *IEEE Transactions on Pattern Analysis and Machine Intelligence* 31, 563-569.
- [Silva (2006)] Silva, R., Scheines, R., Glymour, C., Spirtes, P., 2006. Learning the structure of linear latent variable models. *Journal of Machine Learning Research* 7, 191-246.
- [Singh (2004)] Singh, S., Allanach, J., Haiying, T., Pattipati, K., Willett, P., 2004. Stochastic modeling of a terrorist event via the ASAM system. *Proceedings of the IEEE International Conference on Systems, Man & Cybernetics, Hague*, 5673-5678.
- [Watts (1998)] Watts, D. J., Strogatz, S. H., 1998. Collective dynamics of small-world networks. *Nature* 398, 440-442.